

VU Research Portal

Three ways to quantify uncertainty in individually applied "minimal important change" values

de Vet, H.C.W.; Terluin, B.; Knol, D.L.; Roorda, L.D.; Mokkink, L.B.; Ostelo, R.W.J.G.; Hendriks, E.J.; Bouter, L.M.; Terwee, C.B.

published in

Journal of Clinical Epidemiology
2010

DOI (link to publisher)

[10.1016/j.jclinepi.2009.03.011](https://doi.org/10.1016/j.jclinepi.2009.03.011)

document version

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

citation for published version (APA)

de Vet, H. C. W., Terluin, B., Knol, D. L., Roorda, L. D., Mokkink, L. B., Ostelo, R. W. J. G., Hendriks, E. J., Bouter, L. M., & Terwee, C. B. (2010). Three ways to quantify uncertainty in individually applied "minimal important change" values. *Journal of Clinical Epidemiology*, 63(1), 37-45.
<https://doi.org/10.1016/j.jclinepi.2009.03.011>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

Three ways to quantify uncertainty in individually applied “minimally important change” values

Henrica C.W. de Vet^{a,b,*}, Berend Terluin^{a,c}, Dirk L. Knol^{a,b}, Leo D. Roorda^d,
Lidwine B. Mokkink^{a,b}, Raymond W.J.G. Ostelo^{a,b,e}, Erik J.M. Hendriks^f, Lex M. Bouter^g,
Caroline B. Terwee^{a,b}

^aEMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands

^bDepartment of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands

^cDepartment of General Practice, VU University Medical Center, Amsterdam, The Netherlands

^dDepartment of Rehabilitation Medicine and Psychology, Jan van Breemen Institute, Amsterdam, The Netherlands

^eFaculty of Earth and Life Sciences, Institute of Health Sciences, VU University Amsterdam, Amsterdam, The Netherlands

^fDepartment of Epidemiology, Maastricht University, Maastricht, The Netherlands

^gExecutive Board of VU University Amsterdam, Amsterdam, The Netherlands

Accepted 16 March 2009

Abstract

Objective: Determining “minimally important change” (MIC) facilitates the interpretation of change scores on multi-item instruments. This article focuses on how MIC values should be interpreted when applied to individual patients.

Study Design and Setting: The MIC value of a hypothetical questionnaire “Q” was determined in a sample of 400 patients who improved and 100 patients who did not improve, using the receiver operating characteristic (ROC) method, and three methods to quantify the uncertainty.

Results: The MIC value on questionnaire Q was 10.5. Firstly, the 95% confidence interval (CI) of the MIC value (for questionnaire Q: 5.6–14.2) quantifies the uncertainty of the estimation of the MIC value. Secondly, “how sure we are that this MIC value holds for every patient” is quantified by the values for sensitivity (74%) and specificity (91%). Thirdly, the smallest detectable change (SDC) on questionnaire Q is calculated (16.0) to consider whether the MIC value (10.5) falls outside or within the measurement error.

Conclusion: For application in clinical research and practice, MIC values are always considered at the individual level, but determined in groups of patients. The interpretation comes with different forms of uncertainty. To appreciate the uncertainty, knowledge of the underlying distributions of change scores is indispensable. © 2010 Elsevier Inc. All rights reserved.

Keywords: Change scores; Anchor-based methods; Minimally important change; ROC method; Outcome assessment; Patient-reported outcomes

1. Introduction

Patient-reported outcomes, such as health-related quality of life or perceived disability, are increasingly used in research and clinical practice. These patient-reported outcomes are often measured by multi-item questionnaires. In addition to validity, reliability, and responsiveness, it is important that the numerical value of an outcome measure is interpretable. This means that users understand the meaning of the measurement results [1]. In case of multi-item questionnaires, it is not immediately clear how the observed

scores and changes in scores should be interpreted. Assume that we use a 10-item questionnaire to assess physical functioning in a patient with low back pain, with a total score ranging from 0 to 50 points. What does a score of 35 points or a change score of 5 points mean? Therefore, we want to know which change scores on such outcome measures are minimally important. Jaeschke et al. [2] defined minimal clinical important difference (MCID) as “the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management.” This definition appears to refer to changes that patients perceive as beneficial and the consequences thereof for the management of individual patients. The OMERACT group [3] and De Vet et al. [4] have pointed out the distinction between

* Corresponding author. Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. Tel.: +31-20-444-8176; fax: +31-20-444-6775.

E-mail address: hcw.deviet@vumc.nl (H.C.W. de Vet).

What is new?

Determining ‘minimally important change’ (MIC) facilitates the interpretation of changes scores on multi item instruments.

MIC values are usually applied to individual patients in research and clinical practice.

This paper provides three methods to quantify the uncertainty in the application of MIC values to individuals: 95% CI, sensitivity and specificity, and whether the MIC value exceeds the measurement error.

differences *between* individuals (or groups) and changes *within* individuals (or groups). Therefore, we prefer the term minimally important change (MIC) for use in clinical practice, where it concerns changes within patients.

In the literature, there is considerable confusion about the values of MIC for group level and for individual level. This confusion concerns three issues. Firstly, what exactly is meant by group level, and what is individual level? Secondly, it is questioned whether different methods should be used to determine the MIC value for groups and for individuals, as suggested by Wells et al. [5]. Thirdly, there is confusion whether the MIC values that are determined in groups of patients differ from those applied to individual patients. It has been suggested that MIC values will be higher when applied to individual patients than to groups of patients [6], but, interestingly enough, also the other way around [7]. Therefore, we felt the need to explain the difference between MIC values at group level and those at individual level. In addition, we elaborate on how MIC values can be applied and interpreted for individual patients in clinical practice.

2. Minimally important change at group level and individual level

Methods proposed for the assessment of MIC values can be broadly classified as anchor-based and distribution-based methods [8,9]. Anchor-based methods use an external criterion (anchor) to determine what patients (or their clinicians) consider to be the MIC and relate the changes on the measurement instrument to this criterion. Distribution-based methods relate the observed change to some form of sample variability or to the measurement error of the measurement instrument [9]. As distribution-based methods lack information whether the observed changes are minimally important, anchor-based methods are usually preferred to determine the MIC values, whereas distribution-based methods provide supportive evidence [10,11].

Whatever method is applied, MIC values are always *determined* in groups of patients, either in randomized

controlled trials (RCTs) or in other longitudinal studies. However, the fact that the MIC is determined in a group of patients does not say anything about the level on which the MIC is applied. According to Guyatt et al. [1], classification as MIC for group level or individual level depends on the level on which the anchor is chosen. They made a distinction between population-focused and individual-focused approaches. A population (group)-focused approach needs an anchor at group level and an individual-focused approach needs an anchor at individual level. For example, a population-focused approach is needed for the question “How much reduction in body weight results in a minimally important effect on disease incidence or mortality?” In this example, the anchors are disease incidence or mortality at population level. This kind of approach is often used in the public health field. An individual-focused approach requires an anchor at the individual level to determine which changes are minimally important for individual patients. For example, an individual-focused approach is needed for the question “How much reduction in body weight results in a minimally important improvement in physical functioning in an individual patient?” In this case, the patient-perceived increase in physical functioning of each individual patient is assessed and used as anchor. In clinical research and practice, an individual-focused approach mostly applies.

Suppose that in an RCT, for comparing two interventions to treat low back pain, the Roland Disability Questionnaire (RDQ) is chosen as primary outcome measure and a mean difference of 3 points is found between the two treatment groups. Low back researchers reached consensus that the MIC value for the RDQ is 5 points [12]. How do we use MIC values for the interpretation of these results?

First question is whether this concerns MIC at group level or at individual level. RCTs, like this one, are typically aimed to find the best treatments for individual patients. For this reason, MIC at individual level should be considered here. However, a comparison of the difference in group means, with an MIC based on an anchor at individual level, is not the right thing to do. The reason is that underlying a mean difference of 3 points on the RDQ, there is a distribution of individual change scores, for example, before and after treatment, in both trial arms. Some of these changes may be small, and others may be substantial. Some may be clinically unimportant, and others may be clinically important. Guyatt et al. [13] have explained how an MIC value can be used as a response criterion in the analyses of trials. One determines for each patient whether the change is larger or smaller than the MIC value, and calculates for each trial arm the proportion of patients whose scores exceed the MIC value, that is, more than 5 points on the RDQ. Subsequently, one can assess the proportion of success in each trial arm and calculate risk differences and number of patients needed to treat by making use of the MIC as a response criterion. This is the way to apply MIC values at individual level in the interpretation of RCTs. To determine whether *differences*

between mean values are minimally important, it is necessary to have an anchor at population level.

After this explanation of the difference between MIC values at group level and individual level, the remainder of the article focuses on how MIC values can be applied and interpreted for individual patients in clinical practice. For this purpose, we present a fictive example. The advantage of a fictive example over a real example is that full attention can be paid to specific methods to express uncertainty around the MIC values, not confused and distracted by comments about the practicalities of a specific example. Finally, in the Discussion section, we will elaborate on the additional challenges that occur with a real example, and reflect on the methods that can be used to interpret MIC values for individual patients.

3. Intermezzo: the fictive example

3.1. Methods

Suppose that, in a longitudinal study, 500 patients with low back pain complete the hypothetical questionnaire Q on physical functioning on two occasions: before and after treatment. From these patients, we have information on an anchor as to whether their physical functioning remained stable or showed slight or much improvement or deterioration. The anchor is considered to be a perfect gold standard in this example. Questionnaire Q is a 10-item questionnaire, and each item is scored on a 6-point scale ranging from 0 (unable) to 5 (no trouble). The total score, obtained by summing the item scores, ranges from 0 (very poor physical functioning) to 50 (perfect physical functioning). Table 1 shows the mean change scores and standard deviations (SDs) of questionnaire Q for the categories on the anchor. The Spearman correlation coefficient between the anchor and the change scores on the measurement instrument is 0.67. We generated the data by assuming a normal distribution of change scores of questionnaire Q within

each category. For example, the 200 patients in the category “much improved” according to the anchor were assigned values by SPSS, leading to a mean value of 20 and an SD of 15. Both the intended and the resulting means and SDs are presented in Table 1.

3.1.1. Data analysis

To estimate the MIC value, we used the receiver operating characteristic (ROC) method [14], visualized by the anchor-based MIC distribution [15]. The ROC method is an anchor-based method, which draws a parallel with diagnostic studies [14]. The change score on questionnaire Q is the diagnostic test result, diagnosing the occurrence of important improvement. The anchor is used as the reference standard. On our anchor, we have categories of patients who are completely recovered, much improved, slightly improved, unchanged in health status, slightly deteriorated, much deteriorated, and worse than ever. In this fictive example, there were no patients with any substantial deterioration in their physical functioning. We decided to label patients who reported on the anchor to be slightly improved, much improved, or completely recovered as “importantly improved,” and patients who reported to be not changed or slightly deteriorated as “not importantly improved.” The distributions of the change scores of the “importantly improved” patients and the “not importantly improved” patients are visualized in a graph representing the anchor-based MIC distribution [15]. The MIC value corresponds to the optimal ROC cutoff point, that is, the value for which the sum of the percentages of false-positive and false-negative classifications ($[1 - \text{sensitivity}] + [1 - \text{specificity}]$) is smallest. The MIC value is based on the empirical ROC curve. Its 95% confidence interval (CI) was obtained by nonparametric bootstrapping (bias corrected and accelerated percentile method) [16,17]. Bootstrapping yields an empirical distribution of the optimal cutoff score, and the boundaries of the 95% CI are given as the 2.5% and 97.5% percentiles.

3.2. Results

3.2.1. Determining the minimally important change value

The anchor-based MIC distribution for questionnaire Q is depicted in Fig. 1. The distribution of the change scores of the 400 patients who showed “important improvement” according to the anchor is presented on the left-hand side, and the distribution of the 100 patients who showed “no important improvement” according to the anchor is presented on the right-hand side. Note that for the estimation of the ROC cutoff point, the distributions of both curves have equal sizes, that is, it is not the absolute number of patients that is depicted but the fraction of patients (relative frequencies). In addition, note that, to form the group of “importantly improved” patients, not only the category of “slightly improved” patients is used, but also the categories of “much improved” and “completely recovered” patients are added, because the category of “much improved” patients might include patients

Table 1
Mean values and standard deviations (SDs) of the questionnaire Q for each category of the anchor, as intended to generate data, and the results

Anchor	Number of patients	Questionnaire Q	
		Intended mean (SD)	Resulting mean (SD)
Completely recovered	100	30 (15)	31.50 (15.45)
Much improved	200	20 (15)	19.73 (14.55)
Slightly improved	100	10 (10)	10.25 (9.45)
Not changed	75	0 (10)	−0.47 (9.73)
Slightly deteriorated	25	−10 (10)	−12.00 (8.42)
Much deteriorated	0	0	0
Worse than ever	0	0	0
Importantly improved group	400	20	20.30 (15.50)
Not importantly improved group	100	−2.5	−3.35 (10.60)

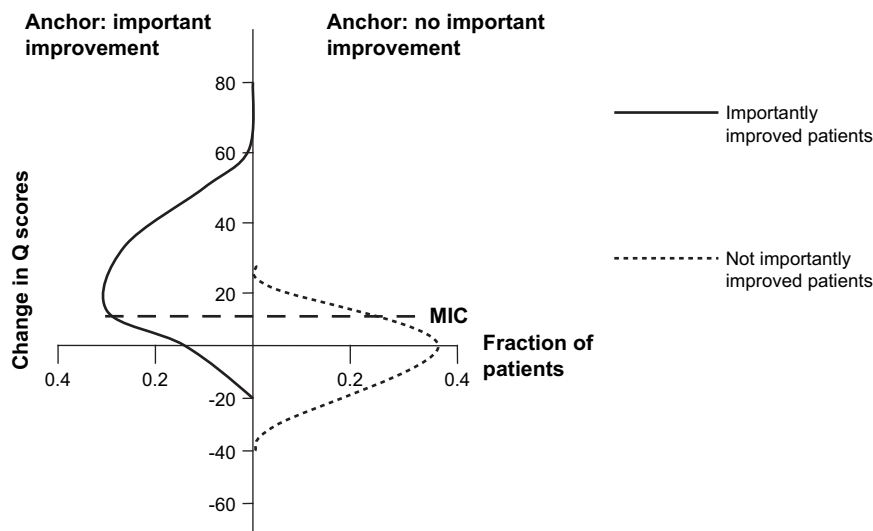


Fig. 1. Anchor-based minimally important change (MIC) distribution of the questionnaire Q with indication of the receiver operating characteristic cutoff point at MIC = 10.5. At this point, sensitivity = 0.74 and specificity = 0.91. For each curve, the relative frequency is presented. Solid curve represents importantly improved patients and dotted curve represents not importantly improved patients.

who score lower than the optimal MIC cutoff point. One certainly wants to label these as false negatives. By the same token, the category of patients who showed deterioration is included in the group of “not importantly improved” patients. The optimal ROC cutoff point lies at a change score on questionnaire Q of 10.5 points. In that case, 26% of the patients who are “importantly improved” according to the anchor are misclassified as “not importantly improved” (sensitivity = 74%), and 9% of the patients who are “not importantly improved” according to the anchor have a higher change score, and are misclassified as “importantly improved” (specificity = 91%). Thus, the MIC value on questionnaire Q is 10.5 points. Now that we have determined the MIC value in this group of patients, we will explain the application and interpretation at individual level.

3.2.2. Application of minimally important change at individual level: an analogy with diagnostic tests

To identify the best cutoff point for a diagnostic test in clinical research, the results of the reference test and the diagnostic test under study are examined in groups of patients. The cutoff point is assessed by minimalizing or finding an optimal trade-off of the percentages of false-positively and false-negatively classified patients. This optimal cutoff point found in research on groups of patients is applied to individual patients in clinical practice. Sensitivity and specificity, and positive and negative misclassifications are expressed as percentages in groups of patients, and translated into probabilities for interpretation at individual level. Hence, going from groups of patients to individual patients in diagnostic studies, the cutoff points remain the same, and the interpretation shifts from proportions to probabilities.

The same reasoning applies to the MIC value determined by the ROC method. Hence, MIC values determined in a diagnostic approach will remain the same when applied

to individual patients. The challenge lies in the interpretation of change scores in individuals even when also here the interpretation shifts from proportions of misclassifications in groups of patients to probabilities in individual patients. Note that, for determining cutoff points in diagnostic tests and for determining MIC values (both at individual level), groups of patients are needed. However, this is different from determining MIC at group level, because in that case, a population-focused anchor is needed.

3.2.3. Interpretation of minimally important change values at individual level

In clinical practice, one has to interpret the change score of an individual patient in relation to the MIC value. The clinician wants to know the probability that, for example, after treatment, a patient has importantly improved, given the observed change score. Suppose that a patient has a change score of 10.5 points, that is, exactly the same value as the MIC. In that case, the probability of this change score in the group of patients who have importantly improved is exactly the same as that in the group of patients who have not importantly improved. A probability of a change score smaller than the MIC value, for example, 5, is greater in the group of “not importantly improved” patients than in the group of “importantly improved” patients, and a probability of a change score larger than the MIC value, for example, 15, is greater in the “not importantly improved” group than in the “importantly improved” group. This is clearly illustrated in the anchor-based MIC distribution (Fig. 1). The probability of a change score of −20 is very unlikely in the “importantly improved” group, and a change score of more than 30 points is highly unlikely in the group of “not importantly improved” patients. How certain one needs to be, greatly depends on the consequences in patient care.

3.2.4. Uncertainties of minimally important change values at individual level

The interpretation of MIC values at individual level is accompanied by uncertainties. An important question therefore is: how confident are we about the MIC values? The uncertainty or confidence can be expressed in three ways:

- A. What is the 95% CI around the MIC value?
- B. How sure are we that the MIC value applies to every patient?
- C. Does the MIC value exceed the measurement error (i.e., is a change equal to the MIC value statistically significant?)

Re A: what is the 95% confidence interval around the minimally important change value? An MIC value obtained in a group of patients should be accompanied with a 95% CI value, which gives an indication of the sampling variation. If the MIC is estimated in a small sample of patients, the 95% CI will be larger, and consequently, the uncertainty whether this is the “true” value increases. The 95% CI for the MIC value was 5.6–14.2.

Re B: how sure are we that the minimally important change value applies to every patient? The answer to the question about how sure we are that the MIC value applies to every patient is illustrated in Fig. 2. On the left-hand side, there is the situation that resembles Fig. 1: the measurement instrument is quite well able to distinguish patients who are importantly improved from patients who are not importantly improved. The MIC value classifies most of the patients correctly. On the right-hand side, there is considerable overlap between the two curves, and although the MIC value may be the same as in the left-hand figure, the percentage of misclassified patients is much higher, as is the range of change scores. This

means that the extent to which the MIC value determined in a group of patients applies to every individual can be expressed by the sensitivity and specificity.

Re C: does the minimally important change value exceed the measurement error? A third interpretation of confidence is the question of whether we are confident that the patient has really changed when he or she has an observed change that is at least equal to the MIC value. In other words, is the magnitude of the MIC value larger than the measurement error? The patients who did not change according to the anchor show an SD of the change scores (SD_{change}) of 9.73 (Table 1). The smallest detectable change (SDC) [18,19], that is, the change beyond measurement error = $1.64 \times SD_{\text{change}} = 16.0$, in which 1.64 represents the z-score corresponding to 95% CI (one-sided). Therefore, values larger than 16.0 points can be considered to exceed the measurement error.

4. Discussion

We will first reflect on the confusion in the literature against the background of the methods we have proposed here. Then, we will justify the use of the ROC method, complemented with the anchor-based MIC distribution, and discuss the practicalities of choosing an adequate anchor and defining MIC on that anchor. We will finish with comments about the uncertainties that accompany the application of MIC values to individual patients.

4.1. Confusion in the literature about minimally important change at group level and individual level

An MIC value at group level can only be obtained by using an anchor at group level [1]. However, when applying

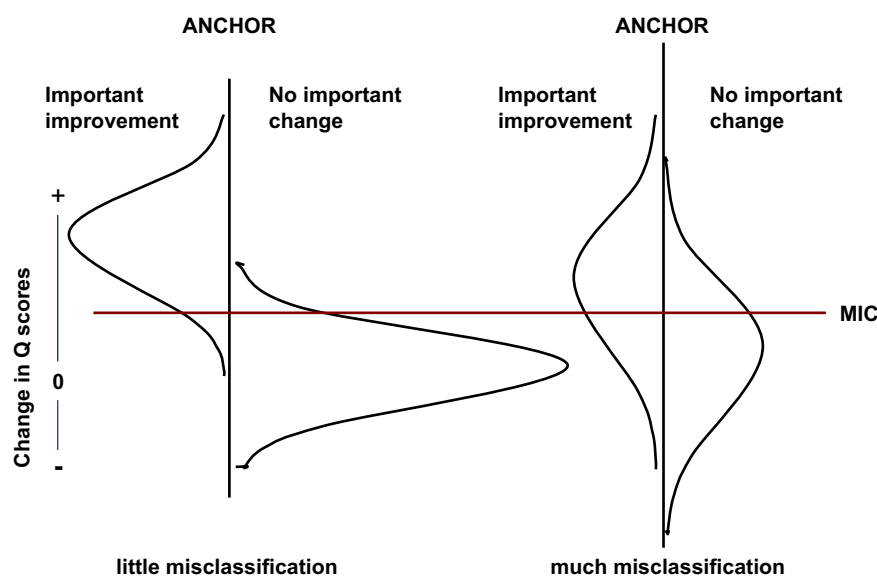


Fig. 2. Two examples of an anchor-based MIC distribution with the same value for the MIC, but different underlying distributions. Left-hand side: little misclassification; right-hand side: much misclassification.

MIC values in RCTs or other longitudinal studies, the anchor is at the individual level. Drawing a parallel with determining and applying cutoff points in diagnostic tests, we have argued that the MIC value as assessed in groups of patients is the best estimate of the MIC value for individual patients. However, some confusion has been created in the literature by authors who have distinguished MIC methods and MIC values for group level and individual level. On behalf of the OMERACT group, Wells et al. [5] classified different methods for assessing MIC in methods at group level and at individual level. Without sound arguments, the ROC method, for example, was labeled as individual level, as was the reliable change index (RCI), whereas the “mean change” method was labeled as “group level” (see the following section for a description and discussion of these methods). However, for both the “mean change” method and the ROC method, the MIC value is determined in groups of patients, and one can subsequently categorize individual patients as below or above this MIC value. Hence, both methods determine the MIC value in groups of patients, and both can be applied in the same manner to individual patients.

With regard to the magnitude of the MIC values, Norquist et al. [6] recommended MIC values which are 1.96 times higher at individual level, that is, for clinicians using information on health-related quality of life to make decisions about individual patients. In contrast, Wyrwich et al. [7] proposed a 50% CI at individual level and 90% or 95% CIs at group level, that is, smaller MIC values for use in individual patients. Hence, both larger and smaller MIC values have been proposed for the individual level. Like others [1,20], we argued that the group and individual levels have the same MIC value, but that more caution is needed for the interpretation at individual level because of accompanying uncertainties.

4.2. Overview of other anchor-based methods to determine the minimally important change value

In the Introduction section, we already stated that the advantage of anchor-based methods is the link with an anchor which defines minimally important change (MIC). However, besides the ROC method, there are other anchor-based methods to estimate the MIC value. A frequently used method is the “mean change” method (adapted from Jaeschke et al. [2]), which takes the mean change score on the measurement instrument for the subcategory of patients who are minimally importantly changed according to the anchor. In our example, the MIC value would have been 10. However, this “mean change” method only provides information about the subcategory of patients who have minimally importantly improved, and gives no information about the change scores of the patients who are not importantly improved according to the anchor. Therefore, it becomes impossible to quantify the probabilities of “importantly improved” vs. “not importantly changed.”

Of course, it is clear that the probability of an important change in patients with a change score above the MIC value will be high, but patients with a score below the mean MIC value may still have a higher probability of belonging to the importantly improved group than to the not importantly improved group. This hampers the interpretation of MIC values obtained by the “mean change” method for individual patients.

Another method is the RCI proposed by Jacobson and Truax [21] to measure individual change. They distinguish clinical relevance and statistical significance. The first step establishes clinical relevance using a cutoff point based on the norm values of a patient/dysfunctional population and a nonpatient/functional population. The second step is intended to investigate statistical significance by assessing whether the RCI is higher than 1.96. The formula for RCI is (observed change)/SD_{diff} [22]. The norm values for nonpatient/functional populations and for patient/dysfunctional populations are often lacking, and this hampers the use of this method. Moreover, it answers the question of whether patients completely recover, in terms of coming in the range of nonpatients rather than whether they experience an MIC.

Another way of calculating the MIC is based on linear regression analysis in which one determines the MIC by estimating the mean change in the instrument corresponding to the known MIC of a similar instrument (anchor) [23]. This method avoids the dichotomization or classification on the anchor. However, this is only possible when there is already a similar instrument for which the MIC value is known. This hampers a broad applicability. Using this method, a 95% CI can be calculated around the MIC value, but it is difficult to quantify the uncertainty when applied to individual patients: it remains unknown whether the MIC value exceeds the measurement error, and it does not provide the probability of a false classification.

4.3. The anchor-based minimally important change distribution is the preferred method to estimate the minimally important change for application at individual level

We chose to supplement the ROC method with a graphical presentation of the anchor-based MIC distribution, because this method combines the advantages of both the anchor-based and the distribution-based methods [15]. Especially in the application at individual level, the anchor-based MIC distribution shows a number of important issues concerning levels of uncertainty, which remain invisible when applying other methods.

1. One can see in the anchor-based MIC distribution that, in our example, the MIC value of 10.5 points is substantially smaller than the SDC, and thus, lies within the measurement error. This means that an observed change in a patient of 15 points, which is

a value between 10.5 and 16.0 points, although higher than the MIC value, may be a chance finding, that is, lies within the 95% distribution of the “not importantly improved” group. However, such a value is more probable for a patient in the group of “importantly improved” patients than for a patient in the group of “not importantly improved” patients. Therefore, this does not disqualify the use of this MIC value in clinical practice, unless one values false-positive classifications (i.e., statistical significance) much higher than false-negative classifications.

2. The anchor-based MIC distribution also illustrates the narrowness of the curves, and consequently, the overlap of the curves for the “importantly improved” group and the “not importantly improved” group. Both depend largely on the correlation between the scores on the anchor and the measurement instrument. Values over 0.30–0.35 [11] and over 0.5 [24,25] have been proposed. In our opinion, values over 0.5 are necessary to prevent too much overlap in the curves and retain reasonable sensitivity and specificity underlying the MIC value. In a previous study, we found sensitivities and specificities of around 0.70 when the correlation between anchor and measurement instrument was in the range of 0.50–0.60 [26]. In this example, the Spearman correlation coefficient between questionnaire Q and the ordinal scale of the anchor was 0.67.
3. The relevant question in clinical practice is whether a patient has experienced at least an important improvement, given the observed change score. This refers to the predictive value of that change score, which is defined as: given this change score, what is the probability that this patient has importantly improved? From diagnostic studies, we know that this predictive value depends on the prevalence of important improvement. Therefore, for a proper interpretation, the clinician should have an idea about the clinical course of the disease or the effectiveness of the treatment. The predictive value can be derived

from the anchor-based MIC distribution, using absolute numbers of patients to draw the curves (Fig. 3) instead of the relative frequencies shown in Fig. 1. The two curves should then be presented based on the sample sizes of the categories of “not importantly improved” patients ($n = 100$) and “importantly improved” patients ($n = 400$).

4. Until now, we have set the ROC cutoff point where the proportion of negative and positive misclassification is the smallest. However, the consequences of positive and negative misclassification may be different. For example, if the consequence of nonimportant improvement is surgery, and a wait-and-see policy does not harm, raising the cutoff point may be more sensible, leading to a higher MIC value. In this way, one reduces the probability of false-positive values, and increases the certainty that a patient has changed if that value is exceeded. In this example, one might require a statistically significant deterioration before deciding to perform an invasive operation.

4.4. Practicalities in the choice of the anchor and defining “minimally important change” on that anchor

We intentionally used a fictive example, because the choice of the anchor and the cutoff point for the definition of minimal importance often gives rise to much discussion. A global rating provided by patients with regard to changes in (specific aspects of) their health status is often used. Such a global rating scale closely links up with the phrase “perceived beneficial by patients” in the definition of Jaeschke et al. [2]. However, critical remarks have been made about such a transition question; firstly, about its reliability because it consists of only one question [27], and secondly, about the fact that it tends to depend more on the most recent measurement than on the first measurement, which is an indication of recall bias [28]. We want to stress, however, that other anchors can be used. For example, Cella et al. [29] used a number of clinical outcomes, such as tumor response and time to progression, as anchor

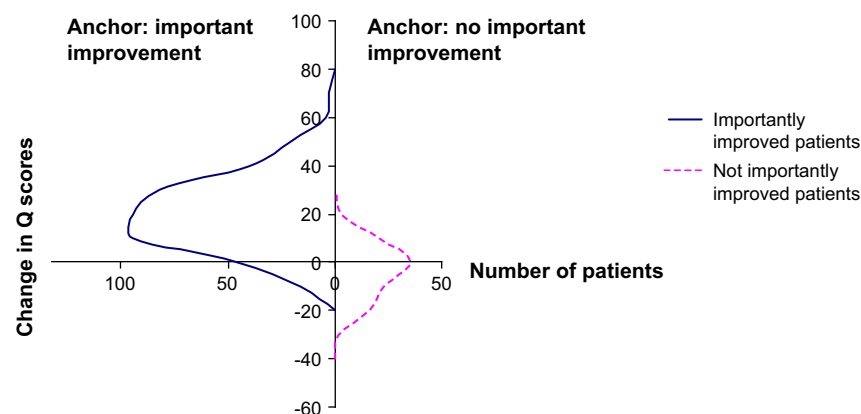


Fig. 3. Distribution of change scores on questionnaire Q in importantly improved and not importantly improved patients. Solid curve represents importantly improved patients and dotted curve represents not importantly improved patients.

in the assessment of the MIC value of a functional assessment questionnaire for lung cancer patients.

The anchors often measure magnitudes of changes, whereas the concept of MIC concerns importance of change. In the literature, studies using a global rating scale have used different cutoff points for defining MIC. Wyrwich et al. have elegantly shown how the definition of the anchor influences the value of MIC [7].

4.5. Comments on the uncertainty around the minimally important change estimations applied to individual patients

A 95% CI gives an indication of the precision of the estimated MIC value in this group of patients. One might question whether it is useful to present these CIs, especially because it is known that MIC values found under different circumstances may differ considerably. It has, for example, been shown that MIC values are dependent on baseline values. With higher baseline values (indicating more severe complaints), the change on the measurement instrument is found to be greater before patients label it as a minimally important improvement on the anchor [30,31]. In addition, it has been found that the MIC values for improvement and deterioration are not always the same [25]. And finally, the choice of the anchor [32] and the type of population [26] influence the resulting MIC value. Thus, it can be argued that presenting narrow CIs around the estimated MIC value might introduce a false sense of confidence. However, as long as one realizes the meaning of a 95% CI, that is, to give insight into the precision in the MIC estimation because of sample variation in *that* particular study, it is informative. Especially with regard to the current recommendation to assess MIC values with different methods, using different anchors [10,11,33], which hopefully triangulate to one MIC value or a small range of MIC values, knowing the 95% CI around the estimations is very useful.

The values of sensitivity and specificity at the optimal ROC point, that is, the chosen MIC value, also have direct bearing on the certainty with which the MIC value can be applied to individual patients: high sensitivity and specificity values minimize the probabilities of misclassification. We advise against the application of the MIC to individual patients when sensitivity and specificity are less than 75%.

Observing whether the MIC values exceed the measurement error also has an impact on the interpretation of MIC in individual patients. If the questionnaire Q is known to have a substantial measurement error, then observed changes in individual patients, although higher than the MIC value, may still be attributed to measurement error. Note that this was the case in our fictive example, where the MIC value of 10.5 did not exceed the SDC. What should we conclude in these situations? The main conclusion is that the measurement instrument is not suitable as a guide in the clinical management of an individual patient. If it is used, nevertheless, one has a high chance of measurement error.

5. Conclusion

The MIC values determined in groups of patients can be applied to individual patients. The ROC method, complemented by a graph of the anchor-based MIC distribution, provides all the necessary information for the quantification of the uncertainty when MIC values are applied to individual patients.

References

- [1] Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, the Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371–83.
- [2] Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;10:407–15.
- [3] Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID working group. Outcome measures in rheumatology: minimal clinically important differences. *J Rheumatol* 2001;28:400–5.
- [4] De Vet HCW, Beckerman H, Terwee CB, Terluin B, Bouter LM. Definition of clinical differences. *J Rheumatol* 2006;33:434–5.
- [5] Wells GA, Beaton D, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differences: review of methods. *J Rheumatol* 2001;28:406–12.
- [6] Norquist JM, Fitzpatrick R, Jenkinson C. Health-related quality of life in amyotrophic lateral sclerosis: determining a meaningful deterioration. *Qual Life Res* 2004;13:1409–14.
- [7] Wyrwich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? *J Biopharm Stat* 2004;14:97–110.
- [8] Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res* 1993;2:221–6.
- [9] Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;56:395–407.
- [10] Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD J Chronic Obstructive Pulm Dis* 2005;2:63–7.
- [11] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102–9.
- [12] Ostelo RW, Deyo R, Stratford P, Waddell G, Croft P, Von Korf M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine* 2008;33:90–4.
- [13] Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ* 1998;316:690–3.
- [14] Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change. An analogy to diagnostic test performance. *J Chronic Dis* 1986;39:897–906.
- [15] De Vet HCW, Ostelo RW, Terwee CB, Van der Roer N, Knol DL, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res* 2007;16:131–42.
- [16] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
- [17] Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000;19:1141–64.

- [18] Beckerman H, Roebroeck ME, Lankhorst GJ, Becher J, Bezemer PD, Verbeek ALM. Smallest real difference: a link between reproducibility and responsiveness. *Qual Life Res* 2001;10:571–8.
- [19] De Vet HCW, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006;4:54.
- [20] Cella D, Bullinger M, Scott C, Barofsky I, Clinical Consensus Meeting Group. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clin Proc* 2002;77:384–92.
- [21] Jacobsen NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;59:1153–60.
- [22] Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining clinical significance of treatment effects: description, application and alternatives. *J Consult Clin Psychol* 1991;59:1153–60.
- [23] Schünemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. Evaluation of minimal important difference for the feeling thermometer and the St George's Respiratory Questionnaire in patients with chronic airflow obstruction. *J Clin Epidemiol* 2003;56:1170–6.
- [24] Guyatt GH, Jaeschke RJ. Reassessing quality of life instruments in the evaluation of new drugs. *Pharmacoeconomics* 1997;12:1478–87.
- [25] Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual Life Res* 2002;11:207–21.
- [26] Van der Roer N, Ostelo RW, Bekkering GE, Van Tulder MW, De Vet HC. Minimally clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. *Spine* 2006;31:578–82.
- [27] Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol* 1997;50:869–79.
- [28] Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition rating scales. *J Clin Epidemiol* 2002;55:900–8.
- [29] Cella D, Eton DT, Fairclough DL, Bonomi P, Heyes AE, Silberman C, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. *J Clin Epidemiol* 2002;55:285–95.
- [30] Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part I. *Phys Ther* 1998;78:1186–96.
- [31] Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001;94:149–58.
- [32] Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. *Arthritis Rheum* 2000;43:1478–87.
- [33] Sloan JA, Cella D, Hays RD. Clinical significance of patient reported questionnaire data: another step toward consensus. *J Clin Epidemiol* 2005;58:1217–9.